Towards Facial Expression Recognition in the Wild: A New Database and Deep Recognition System

Xianlin Peng, Zhaoqiang Xia, Lei Li, Xiaoyi Feng
School of Electronics and Information, Northwestern Polytechnical University, China
pengxl515@163.com, zxia@nwpu.edu.cn, li_lei_08@163.com, fengxiao@nwpu.edu.cn

Abstract

Automatic facial expression recognition (FER) plays an important role in many fields. However, most existing FER techniques are devoted to the tasks in the constrained conditions, which are different from actual emotions. To simulate the spontaneous expression, the number of samples in acted databases is usually small, which limits the ability of facial expression classification. In this paper, a novel database for natural facial expression is constructed leveraging the social images and then a deep model is trained based on the naturalistic dataset. An amount of social labeled images are obtained from the image search engines by using specific keywords. The algorithms of junk image cleansing are then utilized to remove the mislabeled images. Based on the collected images, the deep convolutional neural networks are learned to recognize these spontaneous expressions. Experiments show the advantages of the constructed dataset and deep approach.

1. Introduction

Facial expression plays an important role in many fields [9], such as news analysis, image understanding, personalized recommendation, human-computer interface (HCI), etc. Because of this practical importance, automatic facial expression recognition (FER) has attracted the interest of many scholars in the last three decades [24]. However, it is still a challenge that facial expressions are usually classified into limited types while features of each type vary greatly with the race, culture, personality, etc. To classify expressions into limited expression categories accurately, it is necessary for computers to learn various features for each sort of expressions. To achieve this goal, the database for expression feature learning should contain abundant images with various expression features.

Existing databases used for facial expression analysis can be divided into two categories. One class concentrates on the six basic emotions, which are happiness, sadness, surprise, anger, disgust and fear. For example, the MMI facial expression database [15] contains more than 2000 expression images or frames and more than 500 expression images of 50 persons. The Japanese Female Facial Expression (JAFFE) database [4] has 213 expression images of 10 Japanese females with each object posed 3 or 4 examples for each of the seven basic expressions (six emotional expressions plus neutral face). The other class focuses on extracting fine-grained description for facial expressions. For example, the Cohn-Kanade database [5] contains single action units and combinations of action units based on the Facial action coding system (FACS) proposed by Ekman. Besides, there are some other databases that include several common facial expressions, such as neutral [11, 1, 19, 12], wink [1, 19], sleepy [1], talk [19], scream [12]. All the above databases are challenged by two problems: First, the number of images in each database is small; Second, expression images only represent the acted expressions replacing of the spontaneous ones since they were acquired in a highly-controlled environment. As a result, it is difficult to learn abundant expression features and classify expressions effectively.

Most FER approaches work well in the well-controlled databases but are usually invalid for the real-world expression recognition. That is because these constrained datasets have different backgrounds and cannot present the common expressions. For instance, Shan *et al.* [18] have performed the across-dataset experiments, i.e., they performed the classifier training with the Support Vector Machine (SVM) algorithm by extracting LBP feature on the CohnKanade database, and then tested the trained classifiers on the MMI database and the JAFFE database, respectively. They observed that generalization performance across datasets was much lower, such as around 50% on the MMI database and around 40% on the JAFFE database. These results actually reinforce the findings of Littlewort *et al.* [10], in which they trained selected Gabor wavelet features based SVMs on the CohnKanade database and tested them on another database, and obtained the recognition rate of 56%-60%. In order to train a good classifier which can generalize across im-

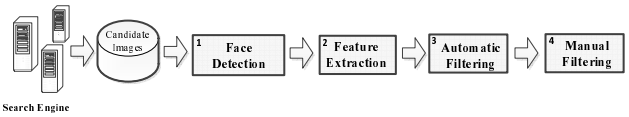

Figure 1. The flowchart of constructing the natural expression database.

age collection environments, the training database should include large-scale expression images to represent various expression features. Few works have been devoted to obtain these expression images. In [22], an interactive construction method was presented for expression image based on the web images. However, the number of expression image for each expression is not sufficient, especially for the learning based algorithms.

In this paper, a novel facial expression database is constructed based on large-scale web images and refers to the different races and cultures. Since these images were labeled by large amount of web users, it is easy to obtain a huge number of images for each kind of facial expressions. Meanwhile, as the web users are not professional persons, many images are usually labeled wrongly or incorrectly. To reach a huge number of effective expression databases, unrelated images (images that were labeled wrongly or incorrectly) should be filtered as junk images. The suggested method includes the following steps: First, large amount of images are obtained by hybrid search engines with query words in different languages, such as happiness, sadness, surprise, anger, disgust and fear respectively, which include expressions images from different people, different background, different race, different age and different environment; then, these returned images are selected initially and clustered by the Affinity Propagation (AP) algorithm; then, fine-grained human labeling is adopted to filter these images for assesses the relevance between the query intentions and search results, and the junk images are filtered interactively. The above steps repeat until the user's requirements are reached.

The rest of this paper is organized as follows. The natural expression database is presented in Section 2 and Section 3 will introduce our deep recognition system. Then we discuss the experimental results for algorithm evaluation in Section 4. At last, Section 5 describes our conclusions.

2. Natural Expression Database

To build up the expression database, six keywords in total (i.e., happiness, sadness, surprise, anger, disgust and fear) are used as query words and search the candidates by the Google and Baidu search engines. Large amounts of images can be returned with each keyword by search engines. The first 2000 images for each expression are chosen for further processing. Then, the 2000 expression images are divided into two groups according to image similarity, which are strongly related or weakly related to one fixed expression. The processing is constituted of four steps shown in Fig. 1, which are introduced in this section.

2.1. Face Detection

Currently, the Viola-Jones face detection algorithm shows the good performance (accuracy and speed) compared to other algorithms [21]. So, in this paper, we utilize it to detect faces from candidates of the original database with the Haar-like features and Adaboost classifiers.

2.2. Feature Extraction

Local texture descriptors, originally designed for grayscale images, can be extended to consider joint texture-color information when combining the features extracted from different color channels. We analyze the color-texture of face images using three local descriptors: Local Binary Patterns (LBP), Local Phase Quantization (LPQ) and Binarized Statistical Image Features (BSIF). Considering that these three features perform well in facial expression representation, they are selected as expression features in this context.

1. **LBP.** The LBP descriptor [13] is a highly discriminative grayscale texture descriptor. For each pixel in an image, a binary code is computed by thresholding a circularly symmetric neighborhood with the value of the central pixel. The occurrences of the different binary patterns are collected into histogram to represent the image texture information.

2. **LPQ.** The LPQ feature [14] is a blur robust image descriptor. The LPQ descriptor is based on the insensitivity of the low-frequency phase components to centrally symmetric blur. Therefore, LPQ employs the phase information of short-term Fourier transform, which is locally computed on a window around each pixel of the image. LPQ is computed using four complex low frequencies: $u_0 = (\alpha, 0)$, $u_1 = (\alpha, \alpha)$, $u_2 = (0, \alpha)$, $u_3 = (-\alpha, -\alpha)$ where α is a small scalar frequency ($\alpha << 1$) ensuring the blur is centrally symmetric. Each pixel x of the image is characterized by a vector F_x of complex frequencies:

$$\begin{aligned} F_x =& [Re\{F(x,u_0), F(x,u_1), F(x,u_2), F(x,u_3)\}, \\ & Im\{F(x,u_0), F(x,u_1), F(x,u_2), F(x,u_3)\}], \end{aligned} \tag{1}$$

where $Re\{.\}$ and $Im\{.\}$ denotes the real part and the imaginary part of a complex number. To maximize the information preservation by the quantization process, the coefficients should be statistically independent. Therefore, a decorrelation step based on a whitening

transform is applied in LPQ before the quantization process. Subsequently, the vector of whitened coefficients is quantized via a simple shareholding scheme:

$$q_i = \begin{cases} 0 \text{ if } f'_i < 0 \\ 1 \text{ otherwise} \end{cases} \tag{2}$$

where f'_i is the ith whitened coefficient. Finally, the resulting binary quantized coefficients are represented as integer values in [0-255] as follows:

$$LPQ(x) = \sum_{i=0}^{8} q_i 2^{i-1} \tag{3}$$

3. **BSIF.** The BSIF feature [6] is a binary texture descriptor belonging to the same family of LBP and LPQ. However, instead of using hand-crafted filters, BSIF utilizes a small set of natural images to automatically learn a fixed set of filters. This set of filters is learned based on the statistics of patches taken from the set of training images. Given a patch X of size $l \times l$ pixels and a linear filter W_i of the same size, the filter response s_i is obtained by:

$$s_i = \sum_{u,v} W_i(u,v)X(u,v) = w_i^T x, \tag{4}$$

where w_i and x are vectors containing the pixels of W_i and X, respectively. A series of binary digits b can be obtained by binarizing each response s_i as follows:

$$b_i = \begin{cases} 1, & if \quad s_i \geq 0 \\ 0, & otherwise \end{cases} \tag{5}$$

where b_i is the ith element of b.

In order to learn a useful set of n filters W_i, the statistical independence of the responses s_i should be maximized. To achieve that, a sufficient numbers of patches are randomly sampled from the training images. The patches are then normalized to zero mean and principal component analysis (PCA) is applied to minimize their dimension to n. Finally, the filters are derived by applying the independent component analysis (ICA) algorithm. Once the filter matrix W is computed, it can be directly utilized for calculating BSIF features from any image.

These three high-dimensional visual features are first extracted and mapped into different feature spaces. Each feature space is used to characterize a certain type of visual properties of images. Kernel can be used to compute the similarity of high-dimensional feature space without too much computational complexity. So suitable base kernels for each feature subset are designed to characterize image similarity. Since these features are histogram based features, the base kernel for histograms is adopted based on the χ^2 kernel function. Given two feature vectors *u* and *v* for images *I* and *J*, the χ^2 kernel is defined below

$$\chi^2 = \frac{1}{2} \sum_{i=1}^{D} \frac{(u_i - v_i)^2}{u_i + v_i} \tag{6}$$

D is the dimension of the feature vector. The u_i and v_i is the *i*th dimension of feature vectors *u* and *v*. Then the kernel function $K_c(I, J)$ for *c*th feature histogram is defined as

$$K_c(I, J) = e^{-\chi^2(u,v)/\sigma_c} \tag{7}$$

σ_c is the mean value of the χ^2 distance between all the images pairs.

At last, the kernels for visual features can be approximated by using a linear combination of these base kernels with different weights and these weights can be determined by cross validation. The diverse visual similarities between images are characterized more precisely by:

$$K_h(I, J) = \sum_{l=1}^{\tau} \beta_l K_l(I, J), \quad \sum_{l=1}^{\tau} \beta_l = 1 \tag{8}$$

where τ is the number of base image kernels. $\beta_l \geq 0$ is the weight for the *l*th base image kernel $K_l(I, J)$ and estimated by cross validation. In this paper, τ is set to three because there are three base kernel functions.

2.3. Automatic Image Cleansing

To further filter weak related images of the original database interactively, it is necessary to display images in above initial database and then remove the weak related images directly. Three steps are constituted in the automatic image cleansing.

First, Affinity Propagation (AP) [2] algorithm is adopted to cluster the images into multiple categories. Considering that it is difficult to display about 2000 images clearly and many images are similar, it is better to display several representative images to people for determining weakly related images. So all images are clustered into some subsets with AP method and then several images are selected as representative images for each category.

Then, hyperbolic visualization is adopted here to have a global view of the initial database [22]. In order to help users assess the relevance between user's query intentions and the initial images interactively, hyperbolic visualization is used to display an amount of returned images according to their nonlinear visual similarity contexts.

After that, unrelated images are filtered as junk images. First, images which are regarded as unrelated images are selected by clicking and dragging the mouse on the hyperbolic visualization space; Then, a few of images that are highly

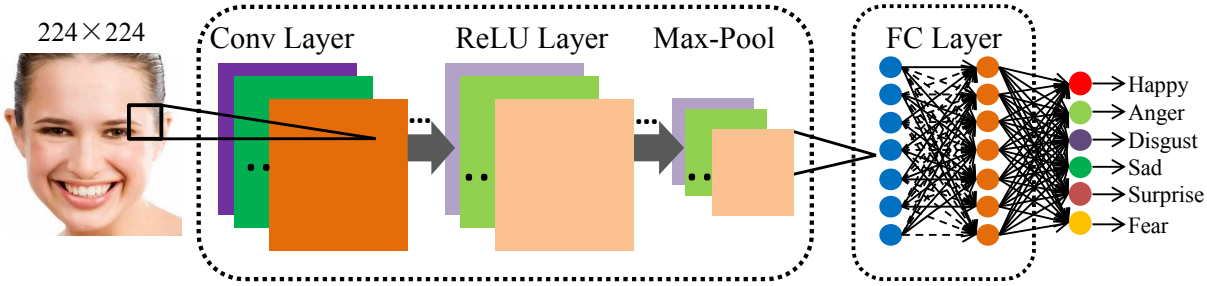

Figure 2. The architecture of our DCNN model.

similar to each selected images are filtered out as junk images.

If it is regarded that there are still many unrelated images, AP clustering is used again to the remaining images and the above interactive filtering procedures are repeated. In this way, junk images are filtered.

2.4. Manual Image Cleansing

After the automatic cleaning by clustering, merely a few images are labeled correctly. Since the amount of these images is relatively small, we choose to eliminate these wrong-labeled images manually.

3. Facial Expression Recognition System

Recently, several deep learning algorithms have been proposed in machine learning and applied to visual object detection and recognition, image classification, face verification and many other research problems [8]. Since several foundational deep learning frameworks, such as Convolutional Neural Networks (CNN) [7], Stacked AutoEncoders (SAE) [20] and Deep Belief Network (DBN) [17], have been presented, numerous deep learning approaches are developed based on these frameworks. These deep learning approaches utilize a large number of images to learn a hierarchy representation and achieve very high performance. The natural expression database can provide sufficient training samples for the model learning.

However, it is still difficult to train the best CNN models by the current facial samples as the CNNs have hundreds of thousands of parameters. Conventional deep models are trained on the data with size of 1 million images. Thus, we propose a new expression recognition method based on pre-trained model in this work. The VGG-face model is initially designed for face recognition [16]. We utilize the deep model to pre-train our model and change the final layer for expression recognition problem.

3.1. Model Architecture

Our deep architecture for expression problem is given in Table 1 consists of 11 blocks, the first 8 blocks are convolutional blocks, where each block contains one or more convolution layers followed by one or more nonlinear layers (ReLU and/or max pooling layers). The last three blocks are fully-connected blocks, the convolutional filters of these blocks have the same size as the input data. This model was pre-trained with a large set of face images (2.6 Million images), its evaluation on two challenging face recognition databases: Labeled Face in the World [3] and Youtube face [23] yield-in to the state-of-the-art results.

To adapt the VGG-face model for our facial expression recognition problem, we fine-tune the model using the constructed natural expression dataset. As in the problem of facial expression recognition, we have six classes (i.e., six basic expressions), we change the output of the last full connected layer from 2622 to 6 (2622 is the number of the subjects used for face recognition). After making these changes, we retrain the deep model by freezing certain layers in the CNN model and learn the weights and biases of the unfrozen layers. To select which layers we should learn we repeat the experiment with different frozen layers. The overall architecture of the proposed method is shown in Figure 2.

In the detailed configuration, the convolutional layer, max pooling layer and fully-connected layer are denoted as the Conv, MPool and FC layer, respectively. The fine-tuning model can extract different texture features of expressions. As shown in Fig. 3, the representation features of natural expressions can be automatically obtained by the deep model, in which only three layers configured in Table 1 have been shown.

4. Experimental Results

4.1. Database

In our experiment, a novel facial expression database with more than 2000 images is established with about 350 images for each kind of expressions. Fig. 4 shows some exemplar images from our expression database for six basic expressions. From the exemplar images, it can be observed that these images involve different races, countries, and ages. Moreover, all images are colored and have different resolutions. Compared with the traditional JAFFE and

Table 1. The configuration parameters in the DCNN.

Layer	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Type	Conv	ReLU	Conv	ReLU	MPool	Conv	ReLU	Conv	ReLU	MPool	Conv	ReLU	Conv	ReLU	Conv	ReLU	MPool	Conv	ReLU	Conv
Support	3	1	3	1	2	3	1	3	1	2	3	1	3	1	3	1	2	3	1	3
Filt dim	3	–	64	–	–	64	–	128	–	–	128	–	256	–	256	–	–	256	–	512
Num filts	64	–	64	–	–	128	–	128	–	–	256	–	256	–	256	–	–	512	–	512
Stride	1	1	1	1	2	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1
Pad	1	0	1	0	0	1	0	1	0	0	1	0	1	0	1	0	0	1	0	1

Layer	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
Type	ReLU	Conv	ReLU	MPool	Conv	ReLU	Conv	ReLU	Conv	ReLU	MPool	Conv	ReLU	Dropout	FC	ReLU	Dropout	FC	Softmax	-
Support	1	3	1	2	3	1	3	1	3	1	2	7	1	–	1	1	–	1	1	–
Filt dim	–	512	–	–	512	–	512	–	512	–	–	512	–	–	4096	–	–	4096	–	–
Num filts	–	512	–	–	512	–	512	–	512	–	–	4096	–	–	4096	–	–	2	–	–
Stride	1	1	1	2	1	1	1	1	1	1	2	1	1	–	1	1	–	1	1	–
Pad	1	0	1	0	0	1	0	1	0	0	1	0	1	–	1	0	–	0	0	0

The 1st layer The 11th layer The 15th layer

Figure 3. Illumination of selected convolutional layers in our deep model.

Figure 4. Exemplar images from our constructed expression database, which contains a total of six expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise).

and Cohn-Kanade databases, we can find obviously that the facial expressions of our database are more spontaneous, and the data quantity is larger than the JAFFE database and Cohn-Kanade database.

4.2. Model Evaluation

To assess the effectiveness of our proposed deep model on the natural expression database, we conducted extensive experiments evaluating the performance of the proposed model and comparing the results against those of using traditional shallow models and controlled datasets.

It is also worth noting that our constructed database is more diverse than the existing controlled databases in Table 2. This clearly demonstrates the natural dataset can yield in much more robust feature learning and classifiers. The main reason might be that, in our dataset, the expressions are obtained from different persons in diverse cultures and countries.

The recognition rates obtained using different methods are shown in Table 3. As one can expect, the performances of traditional LBP and Gabor features are much worse than that of the learned features with our deep model. This confirms that the relative sensitivity of traditional LBP and Gabor features to diverse expressions. Our proposed method outperforms all shallow models.

Table 3. The recognition rate of tag completion based on our constructed dataset, in which the DCNN algorithm is our proposed approach and others are baseline approaches for comparison.

Approach	Expression						
	Anger	Disgust	Fear	Happiness	Sadness	Surprise	Average
Gabor+KNN	0.567	0.408	0.300	0.667	0.533	0.308	0.464
LBP+SVM	0.700	0.400	0.575	0.733	0.633	0.508	0.593
AU+ KNN	**0.721**	0.440	0.564	0.701	0.633	0.617	0.613
DCNN	0.672	**0.754**	**0.755**	**0.983**	**0.908**	**0.936**	**0.834**

Table 2. The recognition rate of tag completion based on our constructed dataset, in which the DCNN algorithm is our proposed approach and others are baseline approaches for comparison.

Approach	Dataset	
	JAFFE	Cohn-Kanade
Gabor+KNN	0.507	0.508
LBP+SVM	0.500	0.500
AU+ KNN	0.521	0.540
DCNN	**0.760**	**0.669**

5. Conclusion

In this paper, a novel database for natural facial expression is constructed leveraging the social images and then a deep model is trained based on the naturalistic dataset. Amounts of social labeled images are obtained from the image search engines by specific keywords. The algorithms of junk image cleansing are utilized to remove the mislabeled images. Based on the images, deep convolutional neural networks are learned to recognize these spontaneous expression. Experiments shown the advantages of the constructed dataset and DCNN.

Acknowledgments

The authors appreciate the reviewers for their extensive and informative comments for the improvement of this manuscript. This work is partly supported by the Fundamental Research Funds of Northwestern Polytechnical University (NO.G2015KY0302) and the National Aerospace Science Foundation of China (No.20131353015).

References

[1] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 19(7):711–720, 2013. 1

[2] B. Frey and D. Delbert. Clustering by passing messages between data points. *Science*, 315(5814):972–6, 2007. 3

[3] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, 2008. 4

[4] M. Kamachi, M. Lyons, and J. Gyoba. The japanese female facial expression (jaffe) database, 1998. http://www.kasrl.org/jaffe.html. 1

[5] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46–53, 2000. 1

[6] J. Kannala and E. Rahtu. BSIF: Binarized statistical image features. In *International Conference on Pattern Recognition (ICPR)*, pages 1363–1366, 2012. 3

[7] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1090–1098, 2012. 4

[8] Y. Lecun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–44, 2015. 4

[9] S. Li and A. Jain. *Handbook of Face Recognition*. Springer Publishing Company, Incorporated, 2nd edition, 2011. 1

[10] G. Littlewort, M. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop*, page 80, 2004. 1

[11] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding facial expressions with gabor wavelets. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 200–205, 1998. 1

[12] A. M. Martinez. The ar face database. *Cvc Technical Report*, 24, 1998. 1

[13] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 24(7):971–987, 2002. 2

[14] J. Ojansivu, V.and Heikkilä. Blur insensitive texture classification using local phase quantization. In *Image and Signal Processing*, volume 5099, pages 236–243, 2008. 2

[15] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–5, 2005. 1

[16] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *British Machine Vision Conference*, 2015. 4

[17] R. Salakhutdinov and G. E. Hintonr. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006. 4

[18] C. Shan, S. Gong, and P. Mcowan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image & Vision Computing*, 27(6):803–816, 2009. 1

[19] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (pie) database. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 46 – 51, 2002. 1

[20] P. Vincent, H. Larochelle, Y. Bengio, and P. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 4

[21] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004. 2

[22] X. Wang, X. Feng, and J. Peng. A novel facial expression database construction method based on web images. In *Proceedings of the Third International Conference on Internet Multimedia Computing and Service*, ICIMCS '11, pages 124–127, New York, NY, USA, 2011. ACM. 2, 3

[23] L. Wolf and I. Hassner, T.and Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 529–534. IEEE, 2011. 4

[24] Z. Zeng, P. Maja, G. Roisman, and T. Huang. A survey of affect recognition methods: audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 31(1):39–58, 2008. 1